Разведанализ данных

В нашем исследовании мы используем усредненные данные по **50 штатам США с 2010 по 2019 год.**

Чтобы изучить экономические условия штата, мы включили такие характеристики как количество совершенных насильственных преступлений, уровень бедности и безработицы, а также доход и реальный ВВП населения.

Данные были собраны с разных ресурсов, так что требовали некоторых преобразований для дальнейшего анализа.

Предварительный просмотр данных:

Show 10 **entries** **Search:**

	State	Vcrime	PR	UR	PIPC	RGDP
1	Alabama	462.25	0.1805	0.0657017690565541	38436.5	193018.02
2	Alaska	702.87	0.1084	0.0676240971098351	55765.2	54741.09
3	Arizona	432.71	0.1673	0.0688650774555567	39264.4	286210
4	Arkansas	506.21	0.1813	0.0558507627847626	38907.7	111633.63
5	California	433.69	0.15	0.0764721540817871	53327.5	2335825.91
6	Colorado	336.26	0.1173	0.0527386225249409	50676.4	306193.02
7	Connecticut	254.19	0.1021	0.0638962813084863	67553.1	243048.54
8	Delaware	525.88	0.1234	0.0566320515487735	47409.6	62762.05
9	Florida	477.64	0.155	0.0650156278893149	44998.3	849951.18
10	Georgia	375.64	0.1696	0.0689790169076298	41273.3	486648.8

Showing 1 to 10 of 50 entries Previous 1 2 3 4 5 Next

Рассмотрим обощающие данные по всем переменным:

Переменная State

```
   Length     Class      Mode
       50 character character
```

Текстовая переменная, обозначающая название штата.

```
 [1] "Alabama"        "Alaska"         "Arizona"        "Arkansas"
 [5] "California"     "Colorado"       "Connecticut"    "Delaware"
 [9] "Florida"        "Georgia"        "Hawaii"         "Idaho"
[13] "Illinois"       "Indiana"        "Iowa"           "Kansas"
[17] "Kentucky"       "Louisiana"      "Maine"          "Maryland"
[21] "Massachusetts"  "Michigan"       "Minnesota"      "Mississippi"
[25] "Missouri"       "Montana"        "Nebraska"       "Nevada"
[29] "New Hampshire"  "New Jersey"     "New Mexico"     "New York"
[33] "North Carolina" "North Dakota"   "Ohio"           "Oklahoma"
[37] "Oregon"         "Pennsylvania"   "Rhode Island"   "South Carolina"
[41] "South Dakota"   "Tennessee"      "Texas"          "Utah"
[45] "Vermont"        "Virginia"       "Washington"     "West Virginia"
[49] "Wisconsin"      "Wyoming"
```

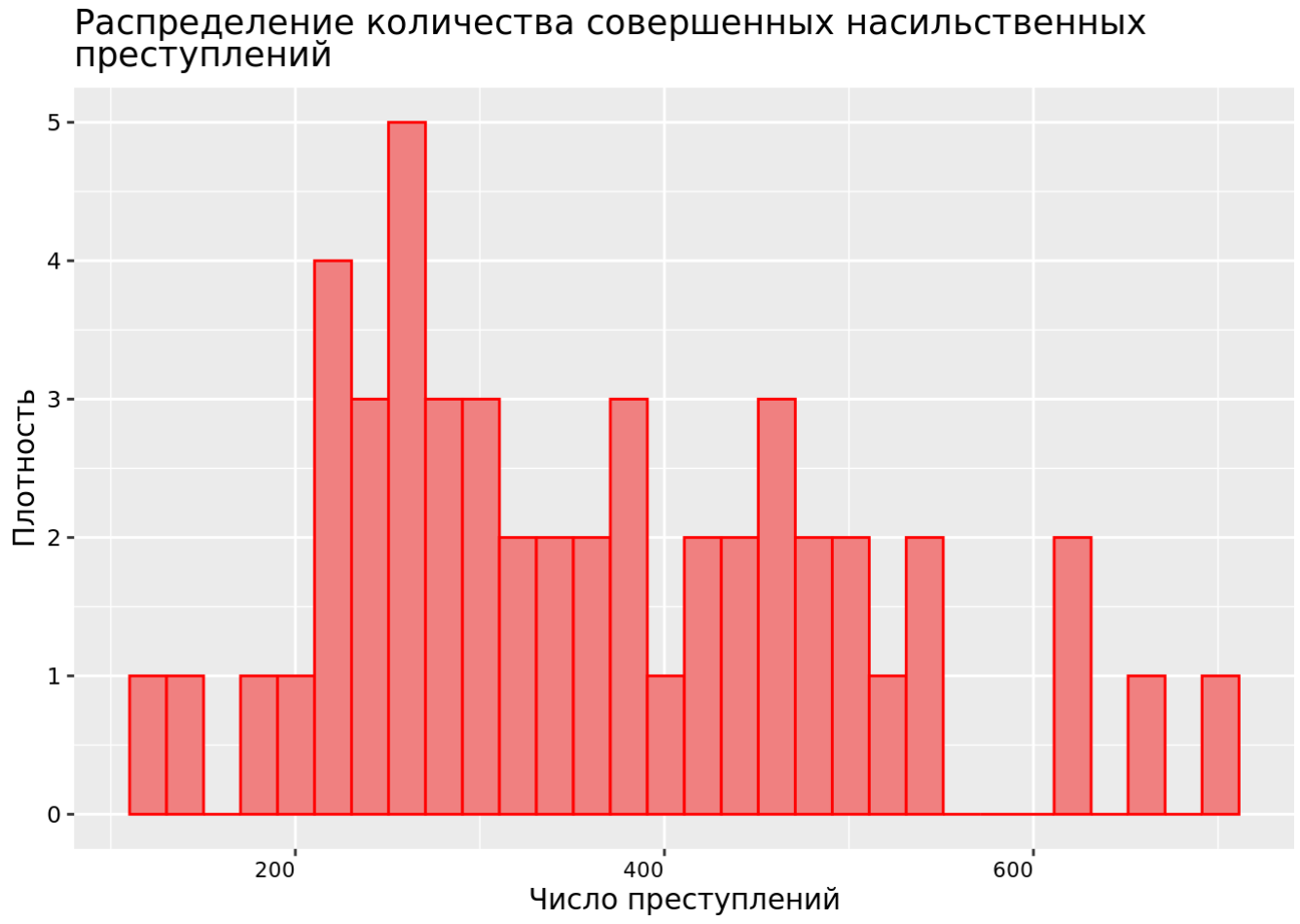
Имеет 50 уникальных значений.

Переменная Vcrime

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  121.7   261.1   337.0   362.9   460.5   702.9
```

Числовая переменная, указывающий количество совершенных насильственных преступлений (violent crimes) на 100 тысяч населения.

Значения колеблются в диапазоне от 122 до 703, где среднее (363) и медианное (337) близки друг к другу.



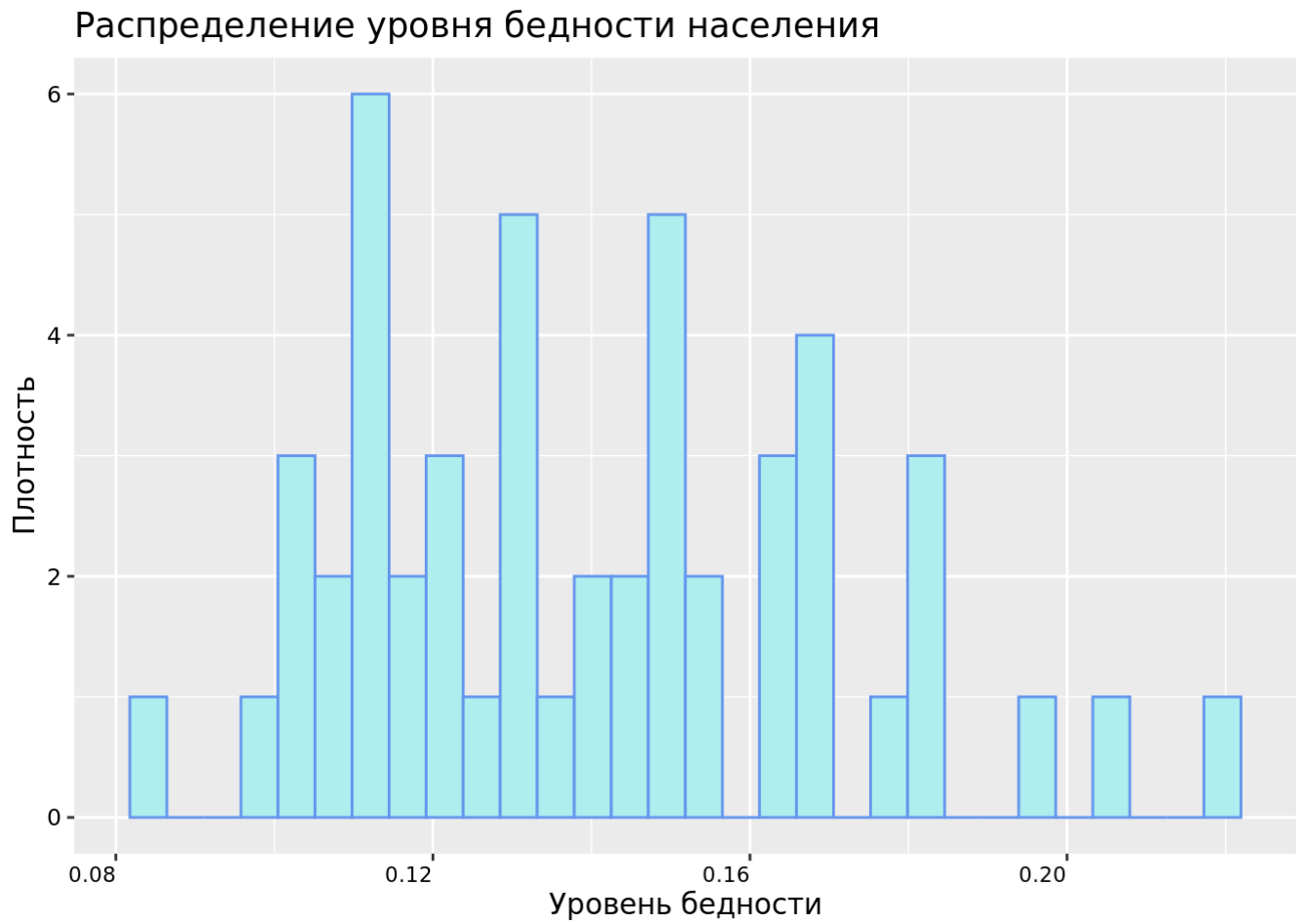
На графике видно, что количество регистрируемых наблюдений в целом находится на достаточно высоком уровне, что еще раз подтверждает выводы СМИ о существующей проблеме с насильственными преступлениями на территории Соединенных Штатов Америки. Статистический службы также зафиксировали и несколько экстремально высоких значений за все время наблюдения (более 600 преступлений). Подобные выбросы можно объяснить снижением затрат на финансирование правоохранительных органов, усилением расслоения общества, экономическими шоками. Следует заметить, что это могло произойти и вследствие ошибок при фиксации и регистрации насильственных преступлений.

Переменная PR

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0818  0.1146  0.1383  0.1405  0.1634  0.2173 
```

Poverty rate - числовая переменная, означающая уровень бедности в долях, то есть от 0 до 1.

В нашем случае минимальное значение 0.0818, а максимальное 0.2173

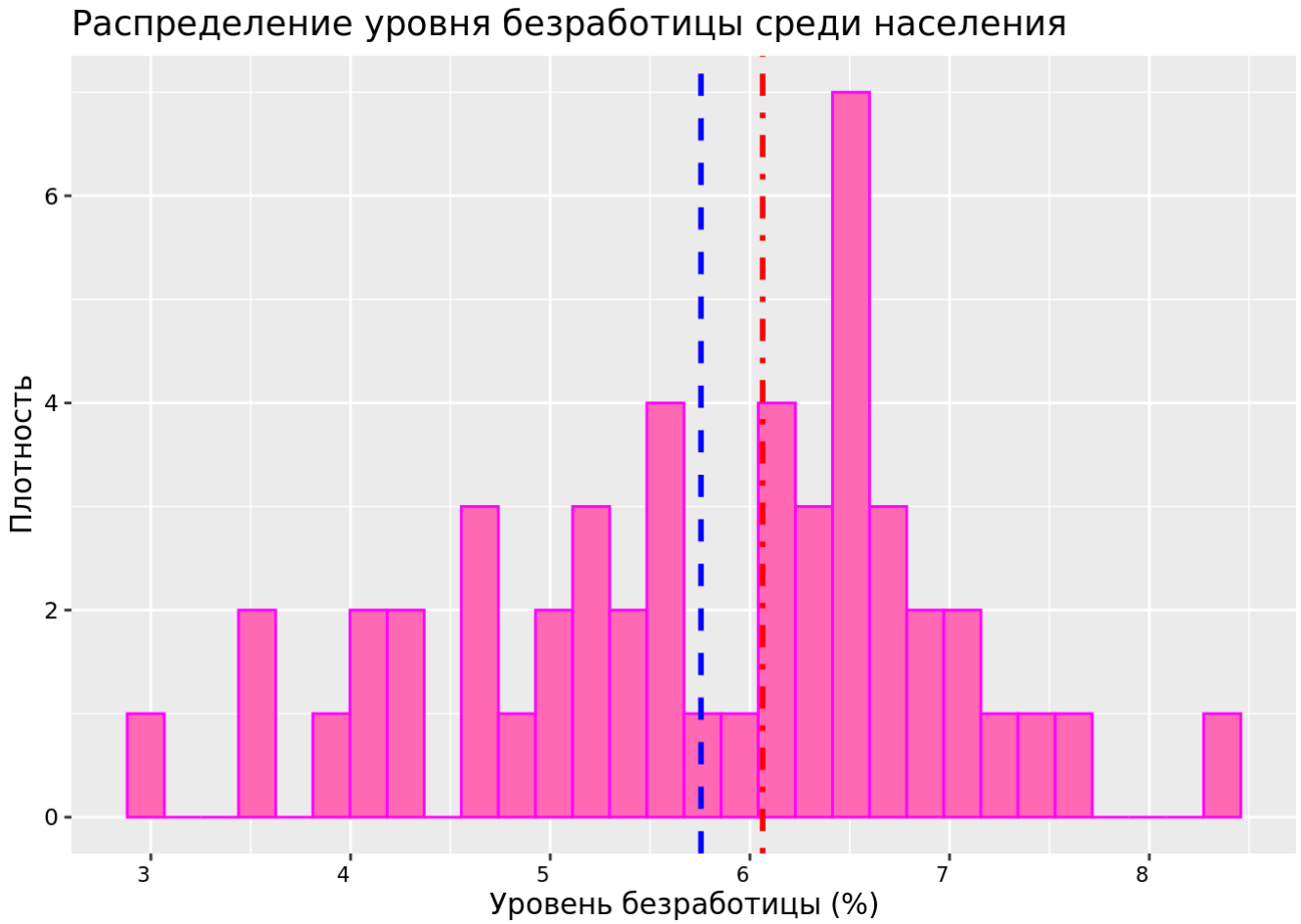


В среднем за весь период наблюдений почти все штаты показывают неплохие - относительно мировых - результаты. Тем не менее, на графике мы можем наблюдать штаты, которые за десять лет наблюдений все так же имеют достаточно высокую (близко к 0,2 и выше) долю населения, находящегося за чертой бедности. Это может объясняться локальными экономическими, географическими и климатическими особенностями, которые мешают развитию штата.

Переменная UR

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.02884 0.04942 0.06064 0.05755 0.06559 0.08275 
```

Unemployment rate - уровень безработицы в штатах в долях. Значения колеблются от 2,9% до 8,3%



На графике видно, что среднее значение и медиана безработицы находятся в пределах 6 %, как было в таблице.

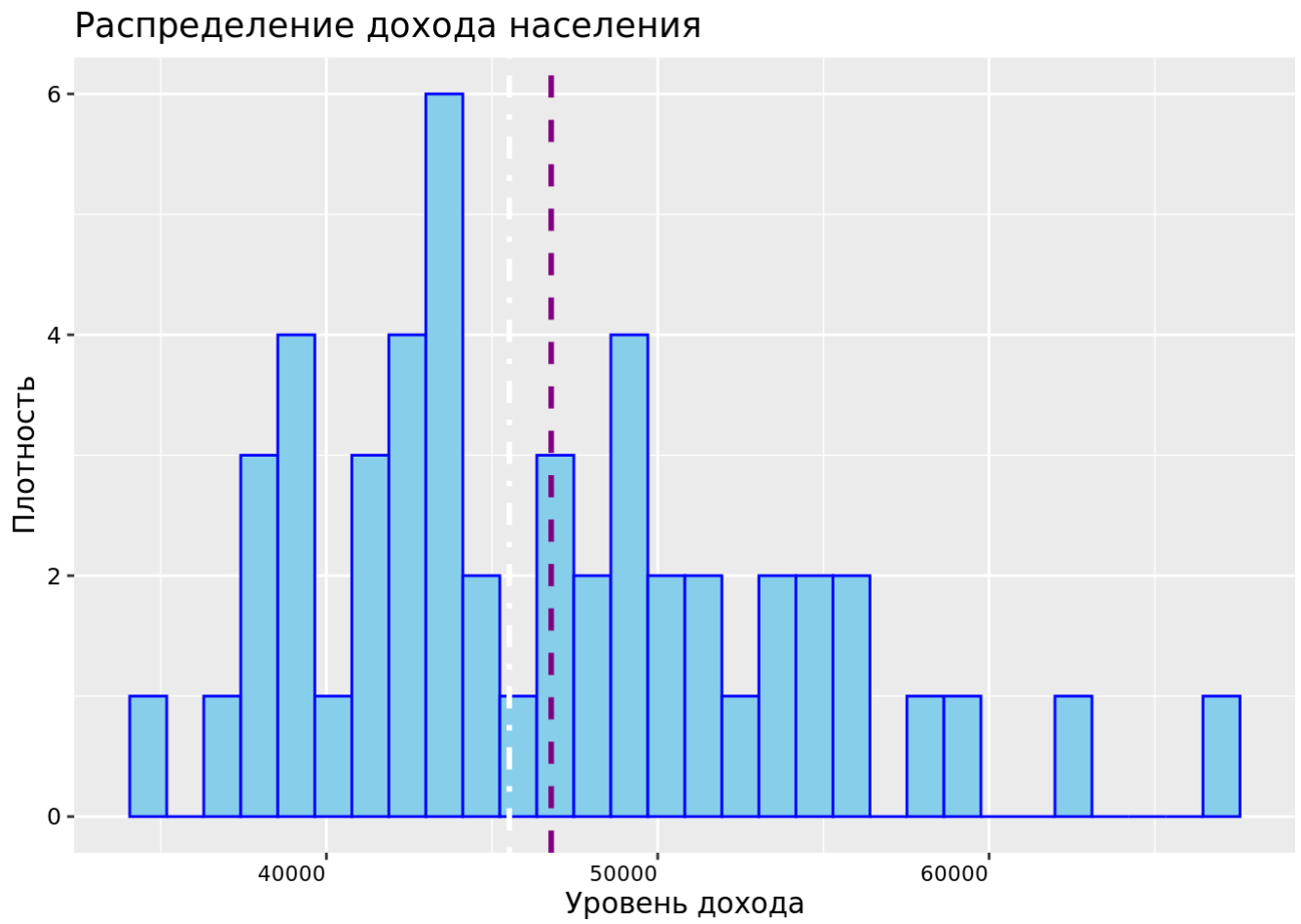
Также видно, что в среднем в Соединенных Штатах Америки сохраняется достаточно низкий уровень безработицы, что может указывать на эффективность работы служб занятости и наличии свободных рабочих мест. Выбросы на графике можно объяснить локальными коллапсами или, например, длительным восстановлением после кризиса 2007-2009 годов.

Переменная PIPC

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  35163   41759   45521   46785   50864   67553 
```

Personal income per capita - числовое значение, выражающее доход на душу населения в долларах.

Максимальное значение (67553) почти в два раза больше минимального (35163), хотя среднее и медианное значения - чуть больше 45000. Получается, что значений, близких к максимальному, довольно мало.

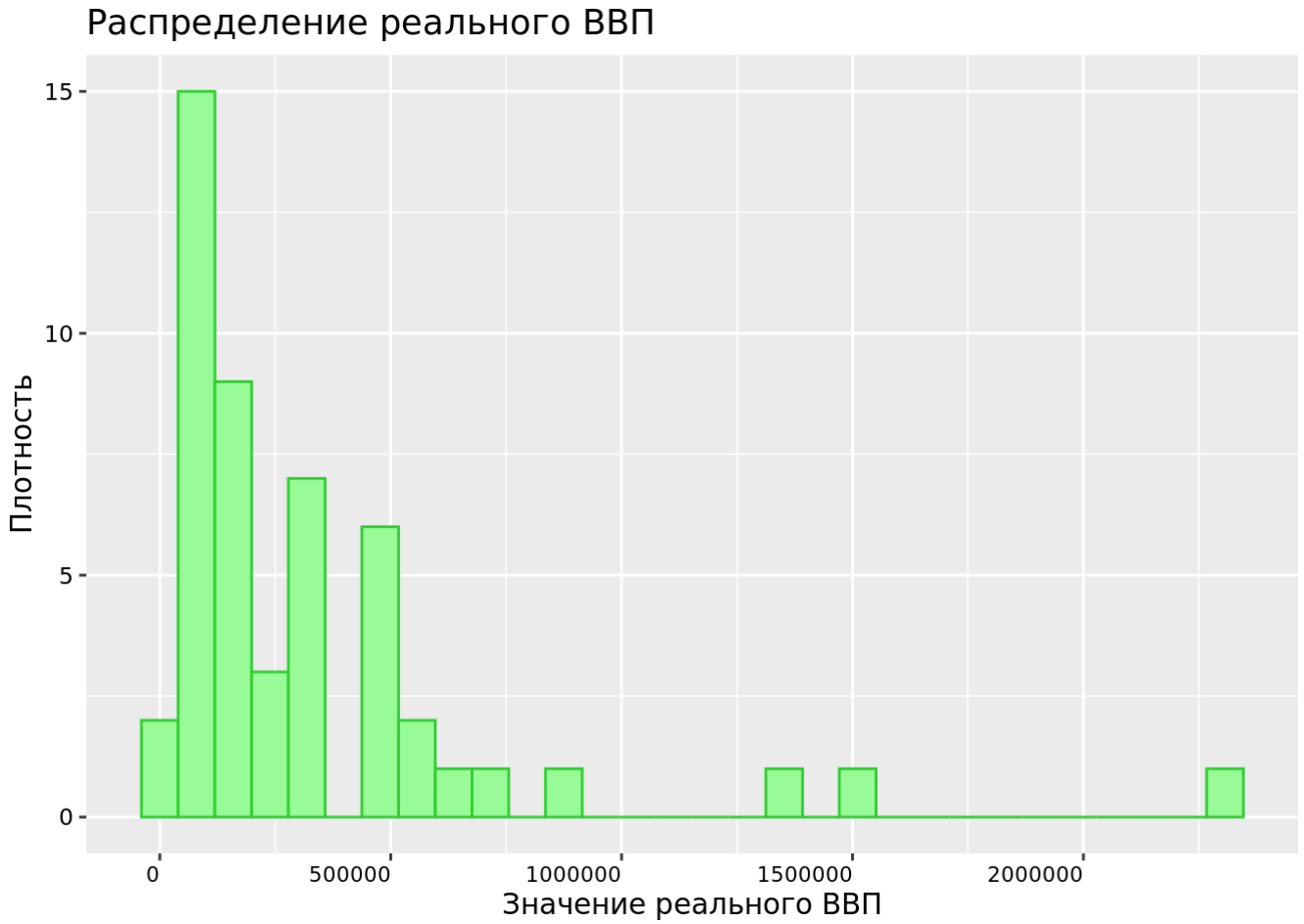


Выбросы на графике можно объяснить спецификой данного показателя: он фиксирует все доходы человека (зарплата, доходы от аренды, дивиденды, трансферты и другие виды). Это приводит к следующей ситуации: в маленьких штатах показатель PIPC может быть либо слишком высоким, либо слишком низким из-за небывалой урожайности (особенно актуально для аграрно-зависимых штатов), природной катастрофы или крупного экономического проекта. В то же время дополнительное население (например, студенты колледжей из других штатов) может занижать показатели PIPC.

Переменная RGDP

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  29180   78561  191783  339150  442790 2335826 
```

Real GDP - числовое значение реального ВВП штата. За 10 лет он колеблется от 29180 до 2335826.



По графику видно, что некоторые значения сильно выбиваются по значению ВВП в большую сторону.

Чтобы понять, как именно выглядит выброс, отсортируем данные по уровню ВВП в штате:

Show 10 **entries** **Search:**

	State	RGDP
1	California	2335825.91
2	Texas	1550373.9
3	New York	1371720.34
4	Florida	849951.18
5	Illinois	741158.3
6	Pennsylvania	672351.99
7	Ohio	570488.87
8	New Jersey	529857.14
9	Georgia	486648.8
10	North Carolina	471399.62

Showing 1 to 10 of 50 entries

Previous 1 2 3 4 5 Next

Таким образом, мы видим, что штат Калифорния отличается от других штатов тем, что средний реальный ВВП с 2010 по 2019 год был выше 2000000, а у остальных штатов значительно ниже. Нельзя назвать это выбросом данных, так как все значения принадлежат одному штату за все года, то есть это реально возможные значения.

Нетрудно объяснить подобное отклонение реального ВВП: Калифорния - крупнейшая экономика в рамках Соединенных Штатов Америки, пятая по показателям ВВП даже среди стран. Одними из драйверов подобного успеха являются численность населения и диверсификация экономики (финансовый сектор и недвижимость, информационная отрасль, производственный сектор вносят наибольший вклад).

Общий вывод

Нам удалось собрать датасет с усредненными данными за обширный период времени, на основе которого мы сможем проверить нашу гипотезу о наличии зависимости между экономическими условиями штата и количеством насильственных преступлений в нем. В рамках разведанализа нам удалось провести чистку данных и убедиться в правильности выбора переменных в рамках нашего исследования.